João Carlos Fidalgo Pinho Oliveira

Imputação em *datasets* médicos – uma comparação entre três métodos

Universidade de Aveiro Departamento de Matemática
2018

João Carlos Fidalgo Pinho Oliveira

Imputação em *datasets* médicos – uma comparação entre três métodos

Relatório de estágio apresentado à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Matemática e Aplicações, realizada sob a orientação científica do Doutor Luís Miguel Almeida da Silva, Professor Convidado do Departamento de Matemática da Universidade de Aveiro e do dr. Bernardo Marques da empresa Prologica.

o júri

presidente

Doutor Agostinho Miguel Mendes Agra

Professor Auxiliar do Departamento de Matemática da Universidade de Aveiro

vogais

Doutor Bruno Miguel Alves Fernandes do Gago

Professor Auxiliar Convidado do Departamento de Ciências Médicas da Universidade de Aveiro

Doutor Luís Miguel Almeida da Silva

Professor Auxiliar Convidado do Departamento de Matemática da Universidade de Aveiro (orientador)

agradecimentos

Agradeço primeiramente à empresa Prologica pela oportunidade proporcionada e pelo conhecimento transmitido ao longo de todo o estágio. Agradeço também a todos os seus membros que sempre estiveram disponíveis para ajudar e que me acolheram na sua família. Agradeço às minhas colegas de estágio Daniela Moleiro e Joana Ferreira com as quais foi um prazer estagiar e das quais guardo muito boas memórias.

Agradeço ao Professor Luís Silva pelo conhecimento e apoio que me transmitiu, bem como pela paciência que teve para comigo ao longo destes meses. Agradeço de igual forma ao Bernardo Marques por todas as ferramentas, conhecimento e disponibilidade que me forneceu durante o estágio.

Agradeço também à minha família e amigos que sempre me deram forças para continuar no meu percurso, agradecendo de forma especial aos meus pais por todos os sacrifícios que fizeram para que concluísse esta etapa da minha vida, e ao meu sobrinho Miguel por me alegrar todos os dias, especialmente quando mais precisava.

Agradeço por fim à Joana Dias, a minha companheira de todos os momentos, que sempre me apoiou e que me ajudou a ser uma melhor pessoa ao longo de todos estes anos. Obrigado por me teres acompanhado desde o início e por sempre teres acreditado em mim.

Palavras-chave

Valores omissos, análise de dados, imputação, MICE, kNN, missForest, regressão, classificação

Resumo

Nos dias de hoje existe um grande volume de dados disponíveis e inúmeros algoritmos que permitem analisar estes conjuntos. No entanto, a maioria dos algoritmos necessita que o conjunto de dados seja completo, isto é, não pode possuir valores omissos. Existem então métodos de imputação que permitem fazer o tratamento dos valores omissos. Neste estudo foram comparados três métodos disponíveis no *software* R, comparando a sua performance em conjuntos de dados na área da saúde disponíveis no UCI *Machine Learning Repository*, com tipos de variáveis mistas (numéricas e categóricas). Foram gerados valores omissos para cada conjunto, nas percentagens de 10%, 20%, 30%, 40% e 50%, posteriormente sujeitos a métodos de imputação simples e múltipla. Foram analisados depois os erros de imputação para as variáveis numéricas e categóricas, comparando também o tempo que cada método demorou a imputar cada conjunto de dados, e o seu impacto na classificação. Os resultados mostraram que o método mais consistente a imputar conjuntos de dados clínicos é o missForest, apresentando de forma quase constante o menor erro de imputação, mas devido à sua maior complexidade também é o método que leva mais tempo a imputar.

Keywords Missing values, data analysis, imputation, MICE, kNN, missForest, regression, classification

Abstract

Nowadays there is a great volume of available data and countless algorithms that allows us to analyse it. However, most algorithms only work with complete datasets, with no missing values. To solve this problem there are imputation methods that treat the missing data. In this study three methods available in R were used, comparing their performance in imputing medical datasets available at the UCI Machine Learning Repository, with mixed type variables (numeric and categorical). Missing values were generated for each dataset, creating new datasets with 10%, 20%, 30%, 40% and 50% of missing values, and single and multiple imputation methods were applied. The imputation erros were analysed for each type of variable, numeric and categorical, also comparing the imputation time, as well as the impact that each imputation has on classifying each dataset. The results show that the missForest method is the most consistent for clinical datasets, usually presenting the smaller imputation error, but because of its complexity it's also the method that takes longer to impute the missing values.

Índice

Lista de Figuras

Lista de Tabelas

1. Introdução

Problema em Estudo

Nos dias de hoje toda a informação sobre os registos clínicos dos pacientes é armazenada, o que potencializa a investigação na área da saúde. No entanto, apesar de existirem grandes quantidades de dados disponíveis a sua qualidade, por vezes, não é a melhor, sendo que um dos maiores problemas que existe é a presença de valores omissos. A falta de valores e/ou de observações num conjunto de dados pode comprometer a sua análise, ou mesmo impedir que esta seja feita, visto que a grande maioria dos algoritmos de previsão necessitam de conjuntos de dados completos (sem valores omissos).

Uma forma de resolver o problema de valores omissos é a imputação, no entanto existem múltiplos métodos disponíveis. Este trabalho tem como objetivo tentar encontrar um método de imputação que permita imputar conjuntos de dados clínicos de uma forma consistente. Os três métodos em teste são: kNN, MICE e *missForest.* Mais detalhes sobre estes métodos podem ser consultados no Capítulo 4 – Metodologias de imputação.

Empresa

Este trabalho foi desenvolvido no âmbito de um estágio curricular do mestrado em Matemática e Aplicações na área de especialização em Estatística e Otimização, realizado na empresa Prologica filiada em São João da Madeira. A Prologica, fundada em 1984, cria e implementa soluções nas áreas da saúde, educação e cidadania, trabalhando em vários continentes como a Europa, África e América Latina.

A vertente *healthcare* da empresa tem equipas de várias áreas como *Data Engineering, Data Science* e *Business Intelligence*, que procuram resolver problemas relacionados com a saúde, trabalhando com múltiplas instituições como hospitais, clínicas e sociedades. Mais informação pode ser consultada em www.prologica.pt.

2. Valores omissos

A existência de valores omissos é um dos maiores e mais comuns problemas que afetam a exploração e transformação de dados, e que por sua vez, poderá ter um grande impacto nos resultados finais obtidos. Apesar de existiram formas de lidar com este problema, é importante perceber que nenhuma é perfeita e que todas têm problemas associados.

As duas formas mais comuns de lidar com valores omissos são: a remoção dos dados com valores omissos, ou a imputação dos mesmos. De forma a escolher qual dos métodos se vai aplicar é necessário primeiro compreender qual a razão de os dados estarem em falta. Existem três tipos de valores omissos:

1. *Missing at Random* (MAR)

Quando os valores estão *missing at random* (ou omissos ao acaso) significa que a propensão para um valor estar em falta não está relacionado com a variável em questão, mas sim com alguns dos dados observados [1]. Por exemplo, num ambiente hospitalar, um paciente sem problemas cardíacos é propenso a ter leituras da pressão arterial em falta. Nesta situação, temos que o valor omisso da pressão arterial não se deve à especificidade da variável, mas sim ao facto que o paciente não possui problemas cardíacos.
Este tipo de valores omissos são algo problemáticos, pois significa que os dados podem ser enviesados, ou seja, espera-se que exista uma diferença entre o que seria o valor real e o que foi estimado pelo método de imputação com base nas outras variáveis.

2. *Missing Completely at Random* (MCAR)

Quando os valores estão *missing completely at random* (ou omissos completamente ao acaso) significa que não existe nenhuma relação entre o valor em falta e o resto dos dados, quer estes existam ou não [1]. Isto significa que os valores omissos não estão relacionados nem com a própria variável nem com nenhuma outra [2]. A ocorrência deste tipo de omissos surge quando, por exemplo, respostas a um questionário são perdidas ao acaso. Aqui a falta

valores nas respostas não está relacionada com nenhuma variável. Desta forma, este tipo de valores omissos são os menos problemáticos, pois significa que os dados não são enviesados.

3. *Missing not at Random* (MNAR)

Quando os valores estão *missing not at random* (ou omissos não ao acaso) significa que a razão para um valor estar em falta advém da própria variável, isto é, da sua especificidade ou sensibilidade. Por exemplo, num questionário sobre saúde um individuo que consuma drogas pode decidir omitir informação relacionada com esse facto [1].

Estes tipos de valores omissos podem ser muito problemáticos, significando que os dados podem ser muito enviesados por existir informação deliberadamente omitida [2]. É necessário ter muita atenção no tratamento destes valores, sendo que nem a remoção nem a imputação tendem a gerar bons resultados.

É de notar que os valores MAR e MNAR, ao contrário dos MCAR, devem ser tratados com mais atenção, pois, como a sua ausência não é aleatória, o facto de não existirem pode ser por si informação relevante para análise [3].

3. Tratamento dos valores omissos

Entre as formas de lidar com valores omissos, a imputação é provavelmente a mais utilizada, sendo um passo sempre muito importante no pré-processamento de dados [4]. Consoante as características dos dados poderão existir métodos mais ou menos adequados à imputação, características estas que podem ser: a dimensão dos dados (nº de observações vs. nº de variáveis), o tipo de variáveis (contínuas ou categóricas), ou mesmo a natureza dos próprios dados (por exemplo, métodos adequados para imputar dados demográficos podem não ser os mais adequados para dados médicos).

Podemos dividir os métodos de imputação em duas grandes categorias, a imputação simples e a imputação múltipla. Ambos os métodos têm vantagens e desvantagens consoante o tipo de dados que se pretende imputar, ficando à decisão do utilizador qual deles usar.

Imputação simples vs. Imputação múltipla

A imputação simples é a mais fácil de realizar e requer menos tempo, tendo como resultado apenas um novo conjunto de dados com um valor onde antes se encontrava omisso. Existem vários métodos de imputação simples como a média, mediana, vizinhos mais próximos, regressão logística, entre outros [5]. Embora a forma como determinam o valor a imputar seja diferente, de forma geral o processo consiste em analisar os dados presentes e procurar o valor mais provável para cada omisso.

Para conjuntos de dados de pouca dimensão ou com pouca percentagem de omissos, esta metodologia é uma ferramenta bastante rápida para fazer o seu tratamento, obtendo resultados geralmente satisfatórios. No entanto, para conjuntos de dados com grandes percentagens de omissos estes métodos podem causar um grande viés (diferença grande entre o valor estimado e o ao valor real) [6], pois os valores omissos são apenas imputados uma só vez de acordo com a regra definida pelo método. Para os conjuntos com grandes percentagens de omissos é recomendado o uso da imputação múltipla.

A imputação múltipla usa vários modelos e faz várias imputações de forma sucessiva criando diversos conjuntos de dados completos [6]. Esta imputação é feita com base nos seguintes passos:

1. Imputar os valores omissos usando um modelo apropriado que possui uma determinada variância. Este passo consiste primeiramente em imputar todos os valores omissos do conjunto através de um método de imputação simples, sendo que depois para cada variável (uma a uma) todos os valores imputados são revertidos a omissos fazendo uma nova previsão para cada valor com base no restante conjunto de dados, que por sua vez está completo devido à imputação simples feita.
2. Repetir o primeiro passo um número de vezes a definir.
3. Fazer uma análise dos vários conjuntos de dados produzidos, sendo que cada estimativa vai diferir da anterior.
4. Cálculo dos erros estandardizados.

Esta metodologia permite que os resultados obtidos sejam menos enviesados que os da imputação simples [6], obtendo por norma também um menor erro de imputação.

Como consequência do aumento da complexidade dos métodos, estes tendem a ser mais pesados computacionalmente, tornando o processo mais demorado, o que pode ser uma desvantagem para o utilizador.

De uma forma geral a grande vantagem dos métodos de imputação simples é a sua rapidez ao tratar valor omissos, com o risco de os resultados da imputação serem enviesados. Sendo que a vantagem dos métodos de imputação múltipla é a tentativa de redução do enviesamento, tendo um impacto no tempo que leva a imputar.

4. Metodologias de imputação

Neste trabalho serão comparadas 3 das metodologias de imputação disponíveis em pacotes do R. Estas metodologias serão: "*missForest*", que é um método não paramétrico de imputação com base em *random forests*, "kNN", que é um método de imputação que preenche um valor omisso com base nos vizinhos mais próximos, e por fim o *package* "MICE", um método de imputação que usa várias metodologias para preencher os valores em falta consoante o tipo de variável.

kNN

O método de imputação simples *kNN* (*k-nearest neighbours* ou k-vizinhos mais próximos) é bastante utilizado, devido à sua fácil compreensão e rapidez de execução. Este método tem por base encontrar os k-vizinhos mais próximos (entradas completas) para um valor omisso, preenchendo depois este valor com a ocorrência mais frequente (caso a variável seja categórica), ou com, por exemplo, a média dos vizinhos (caso a variável seja contínua).

A proximidade dos vizinhos é calculada tendo por base uma medida de distância. Por definição é utilizada a distância euclidiana, existindo, no entanto, outras medidas que podem ser utilizadas, como a Máxima, Manhattan, Canberra, Binária ou Minkowski[8].

Sendo kNN um método que mede distâncias entre variáveis surge o problema de medir a distância com variáveis categóricas. Para este tipo de variáveis o método efetua uma passo intermédio que transformas as variáveis categórica em numéricas seguindo a seguinte lógica: Seja *I* uma variável categórica com *n* categorias, as observações de *I* passam a tomar valores de {1,…, *n*} [9].

Neste trabalho a medida de distância foi a euclidiana por ser aquela utilizada por definição, e por ser também a mais utilizada, embora hajam evidencias que esta pode não ser a mais aconselhada para conjuntos de dados mistos [10].

A distância Euclidiana, D, é dada pela seguinte fórmula:

$$D(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \ldots + (p_n - q_n)^2}$$

Onde p e q são elementos a comparar com n características. [16]

É graças à sua simplicidade e boa acurácia que este método tem sido usado em múltiplos casos reais. kNN tem sido usado com bastante eficácia no tratamento de valores omissos em inquéritos conduzidos pelo *Statistics Canada, U.S. Bureau of Labor Statistics*, e pelo *U.S. Census Bureau* [11].

MICE

A MICE (*Multivariate imputation by chained equations* ou Imputação multivariada por equações encadeadas) é uma metodologia de imputação múltipla, que ao invés da imputação simples, tem em conta a incerteza estatística nas imputações efetuadas [7]. Uma das vantagens de usar MICE é o facto que esta abordagem é muito flexível conseguindo imputar múltiplos tipos de variáveis, contínuas ou categóricas.

MICE passo a passo:

1. Todos os valores omissos são temporariamente preenchidos por imputação simples (ex.: variáveis contínuas são preenchidas com a média).
2. Para uma variável ("x") todos os valores imputados temporariamente são eliminados.
3. É feita uma regressão entre "x" e as outras variáveis presentes no conjunto de dados, sendo que "x" é a variável dependente neste modelo. Este modelo de regressão é gerado da mesma forma que seria gerado fora do contexto da imputação.
4. Os valores omissos de "x" são então preenchidos com as previsões resultantes do modelo de regressão. Quando a variável "x" for usada na imputação das varáveis seguintes, não serão usados os valores imputados primeiramente pela imputação simples, mas sim os agora obtidos pela regressão.
5. Os passos 2-4 são repetidos para todas as variáveis com valores omissos. Assim que todas as variáveis tenham sido imputadas é considerado um fim de um ciclo de imputação.
6. Os passos 2-5 são repetidos para um número de ciclos definido pelo utilizador (por definição são 10), com as imputações sendo atualizadas a cada ciclo. No final destes ciclos as imputações finais são guardadas, tendo como resultado um conjunto de dados sem valores omissos.

missForest

missForest é também um método de imputação múltipla bastante flexível, que consegue imputar diversos tipos de variáveis. Apesar de ser uma forma de imputação que por norma gera bons resultados, perde um pouco no tempo de processamento. *missForest* tem um funcionamento semelhante ao MICE, sendo que a maior diferença é que o modelo de regressão utilizado é sempre *Random Forests (RF)* [12].

missForest passo a passo [12]:

1. Todos os valores omissos são temporariamente preenchidos por imputação simples (ex.: variáveis continuas são preenchidas por definição com a média).
2. As variáveis são ordenadas de forma crescente de número total de omissos.
3. Para uma variável ("x") todos os valores imputados temporariamente são eliminados.
4. É feita uma regressão utilizando *Random Forests* entre "x " e as outras variáveis presentes no conjunto de dados, sendo que "x" é a variável dependente neste modelo.
5. Os valores omissos de "x" são então preenchidos com as previsões resultantes do modelo de regressão (RF). Quando a variável "x" for usada na imputação das varáveis seguintes, não serão usados os valores imputados primeiramente pela imputação simples, mas sim os agora obtidos pela regressão.
6. Os passos 2-4 são repetidos para todas as variáveis com valores omissos. Assim que todas as variáveis tenham sido imputadas é considerado um fim de um ciclo de imputação.
7. Os passos 2-5 são repetidos até que se atinja um critério de paragem. com as imputações sendo atualizadas a cada ciclo. No final destes ciclos as imputações finais são guardadas, tendo como resultado um conjunto de dados sem valores omissos.

O critério de paragem, dado por ε, é alcançado quando a diferença entre o conjunto de dados imputados mais recente e o anterior aumenta, relativamente a ambos os tipos de variáveis. A diferença entre os conjuntos, dada por Δ, é calculada de forma diferente para as variáveis numéricas e categóricas [12].

Para as variáveis numéricas tem-se que:

Seja N o conjunto de variáveis numéricas, X o conjunto de dados da imputação mais recente e Y da anterior, e sabendo que i são os índices das variáveis numéricas, a diferença destas variáveis, Δ_N, é calculada através da seguinte fórmula [12]:

$$\Delta_N = \frac{\sum_{i \in N} (X_i - Y_i)^2}{\sum_{i \in N} X_i^2}$$

Para as variáveis categóricas tem-se que:

Seja C o conjunto de variáveis categóricas, Obs as observações de C, n o número total de observações, X o conjunto de dados da imputação mais recente e Y da anterior, e sabendo que i são os índices das variáveis categóricas, a diferença destas variáveis, Δ_C, é calculada através da seguinte fórmula [12]:

$$\Delta_C = \frac{\sum_{i \in C} \sum_{obs=1}^{n} Obs_{X_i \neq Y_i}}{n_{omissos}}$$

Ou seja, é a percentagem de observações imputadas que diferem de uma imputação para a seguinte.

Desta forma o critério de paragem é alcançado quando $\Delta_N < \varepsilon$ e $\Delta_C < \varepsilon$, o que significa que a diferença entre X e Y é desprezível.

5. Métodos de avaliação dos erros de imputação

De forma a avaliar os erros de imputação, ou seja, as diferenças entre os valores imputados e os valores reais, foram utilizadas duas medidas de *performance*, uma para as variáveis contínuas e outra para as variáveis categóricas. Para as variáveis contínuas foi utilizado a raiz do erro médio quadrado normalizada (*normalized root mean squared error*, NRMSE) [4], definido por:

Seja NRMSE o erro, X o conjunto de dados original e Y o conjunto de dados imputado, temos que:

$$NRMSE = \sqrt{\frac{\Sigma_{i=1}^{n}(X_i - Y_i)^2}{\Sigma_{i=1}^{n}(X_i)^2}}$$

Para as variáveis categóricas o erro foi calculado através da proporção de casos mal imputados, ou seja, a divisão entre número de casos mal imputados e o número total de casos omissos das variáveis categóricas.

Após calculados estes erros, foi feita uma comparação entre os resultados obtidos entre os três métodos de imputação.

6. Experiência e fluxo

De forma a testar qual o método com a melhor consistência na imputação de dados, foram selecionados múltiplos conjuntos de dados médicos disponíveis no UCI *Machine Learning Repository*. Estes conjuntos de dados apresentam vários tipos de variáveis contínuas e categóricas, variando também em tamanho, tanto a nível de observações como ao nível das variáveis.

Após obtidos os dados foi feita uma limpeza de forma a garantir que não existiam valores omissos, pois, para efeitos de teste é necessário que os dados sejam completos.

Para cada conjunto foram gerados valores omissos, fazendo variar a sua percentagem, 10%, 20%, 30%, 40% e 50%. Nesta geração foi deixada intacta uma variável *target* em cada conjunto de dados para posteriormente serem aplicados algoritmos de classificação de forma a avaliar as implicâncias que a imputação tem na taxa de acertos. Estas variáveis foram escolhidas com base na descrição fornecida pelo repositório.

Depois de gerados os novos conjuntos de dados foi aplicada a imputação pelos três métodos descritos anteriormente. Para a aplicação destes métodos é necessário definir alguns parâmetros à priori, o número de vizinhos para kNN e o número de ciclos máximo para MICE e *missForest.* Não existe um número de vizinhos ou ciclos que seja o melhor para todos os conjuntos pelo que se definiu um valor padrão para cada método, sendo que para kNN foram usados os 10 vizinhos mais próximos, para MICE o número de ciclos foi limitado a 10, e para *missForest* o número de árvores foi também limitado a 10.

Com a imputação feita foi calculado o erro da mesma, usando o NRMSE para as variáveis contínuas e a proporção de casos mal imputados para as categóricas.

Este processo foi repetido 100 vezes de forma independente, sendo os valores omissos foram gerados de forma completamente aleatória. A geração de valores omissos de forma aleatória (MCAR) foi feita com o intuito de não criar viés nos dados, embora em situações reais os dados omissos sejam mais comummente do tipo MAR ou mesmo MNAR.

Após a análise aos erros de imputação e sua comparação, foram aplicados algoritmos de classificação a todos os conjuntos de dados, originais e imputados, de forma a avaliar o impacto que a imputação tem na classificação. De forma a avaliar este impacto foi comparada a evolução da taxa de acertos com o aumento da percentagem de omissos.

Para a classificação dos conjuntos de dados foram usados 2 métodos, kNN e *Random Forests* usando *10-fold cross-validation.* Foi usada a biblioteca *caret* disponível no R para a aplicação destes métodos, sendo que o número de vizinhos para kNN e o número de árvores para *Random Forests* foram definidos pelo próprio método, visto que este possui uma função que faz *tuning* automático destes parâmetros.

O método de classificação kNN funciona de igual forma ao de imputação (ver Capítulo 4).

O método de classificação *Random Forests* tem como por base árvores de decisão. Este método constrói o modelo separando o conjunto de dados incial em subconjuntos de menor dimensão de forma iterativa, criando ao mesmo tempo uma árvore de decisão. Este processo tempo como resultado uma "árvore" com nódulos de decisão, que podem conter dois ou mais "ramos". Quando estes nódulos apenas contêm um "ramo" significa que se foi tomada uma decisão.

O algoritmo que gera estas árvores usa entropia (grau de incerteza da informação) para calcular a homogeneidade dos subconjuntos, entropia é 0 caso seja completamente homogéneo ou 1 caso o subconjunto esteja igualmente dividido. [17]

A entropia para apenas um atributo (ex.: "Problemas cardíacos: Sim ou Não") é dada pela seguinte fórmula:

$E(S) = \sum_{i=1}^{c} -p_i log_2 p_i$, onde E é a entropia, S é o atributo, c é o número de opções e p são as probabilidades de cada opção.

Para dois atributos a fórmula da entropia é:

$E(S, X) = \sum_{c \epsilon X} P(c) E(c)$, onde c são as opções de X, P é a probidade de c e E(c) é a fórmula de entropia para um atributo.

Desta forma a entropia é usada para calcular o ganho de informação com a separação de um conjunto em subconjuntos.

Método passo-a-passo [17]:

1. Calcular a entropia da variável alvo.

2. Separar o conjunto de dados com base em diferentes atributos (variáveis) e calcular a entropia para cada "ramo".

3. Calcular o ganho de entropia de cada ramo com base na seguinte fórmula:

$$G(S,X) = E(T) - E(S,X)$$

4. Selecionar o atributo com maior ganho de informação para ser um nódulo de decisão.

5. Repetir este processo para todos os ramos.

6. Quando a entropia for igual a 0 encontrou-se um nódulo de decisão e não se procede a mais separações. Enquanto a entropia for maior que 0 continua-se com as separações.

A diferença entre as árvores de decisão e o método *Random Forests* é que para o último, a cada criação de nódulos de decisão as variáveis são escolhidas de forma aleatória e são geradas várias árvores ao invés de apenas uma. Desta forma obtém-se múltiplas árvores de decisão diferentes o que permite reduzir a variância sem aumentar o viés.

No final todas as árvores são testadas e a que obtiver melhor performance é devolvida pelo modelo.

7. Preparação dos datasets e algoritmo de imputação

De forma a testar a qualidade dos métodos de imputação é necessário utilizar conjuntos de dados completos. Para este estudo foram utilizados 18 conjuntos de dados disponíveis em *UCI Machine Learning Repository*, um *website* que disponibiliza mais de 450 conjuntos de dados de múltiplas áreas, incluindo a área médica, área à qual foi aplicado o estudo. Desta forma foram selecionados 18 conjuntos de dados com características vastamente diferentes, tanto ao nível das variáveis como das observações. Ao nível das variáveis, estas variam entre apenas categóricas e apenas numéricas, variando também a quantidade existente, entre 7 e 36. Ao nível das observações os conjuntos de dados também variam bastante.

Conjunto de dados	**Nº variáveis**			**Nº observações**
	Total	Categóricas	Numéricas	
Acute inflammations	8	7	1	120
Autistic spectrum disorder screening data for adolescents	21	19	2	104
Breast cancer	10	1	9	286
Breast cancer wisconsin (prognostic)	34	2	32	198
Breast cancer wisconsin (diagnostic)	32	1	31	569
Breast tissue	10	1	9	106
Cardiotocography	23	2	21	2126
Contraceptive method choice	10	8	2	1473
Cryotherapy	7	3	4	90
Dermatology	35	34	1	366
Diabetic retinopathy debrecen	20	4	16	1151
Fertility	10	8	2	100
Hepatitis	19	13	6	155
Immunotherapy	8	3	5	90
Liver disorders	7	1	6	345
Pima indians diabetes	8	1	7	768
Thyroid disease	21	15	6	7200
Cervical cancer (risk factors)	36	26	10	858

Tabela 7.1: Descrição dos conjuntos de dados utilizados

A existência desta grande variedade de tipo dados, tanto a nível de observações como de tipo de variáveis, foi propositada, pois o intuito deste estudo é encontrar qual o método mais consistente para a imputação de qualquer tipo de dados médicos.

Como seria de esperar, encontrar 18 datasets com apenas casos completos, ou seja, sem omissos, é uma tarefa difícil, pelo que tiveram que ser selecionados alguns conjuntos de dados incompletos, onde os valores omissos existentes foram excluídos. Esta exclusão foi feita com base em alguns critérios: caso os valores omissos fossem maioritariamente de apenas uma variável esta seria excluída, sendo que depois a exclusão seria feita pelas observações, deixando apenas os *complete cases*, caso os valores omissos estivessem dispersos pelo *dataset* aplicava-se apenas o último método. Para aplicar estes critérios foi desenvolvido um algoritmo que devolvia a percentagem de casos omissos para cada variável ou para cada observação, assim, caso uma observação tivesse uma elevada percentagem de omissos, seria eliminada, sendo a mesma lógica aplicada às observações. Estes dois passos foram feitos alternadamente com o intuito de perder o mínimo de informação possível, até que já não existisse valores omissos no conjunto de dados.

Após esta preparação dos dados, estes passaram a ter apenas casos completos, pelo que se pôde passar a geração de omissos (percentagens fixas de 10, 20, 30, 40 e 50 %) para cada conjunto. Este passo foi realizado no R, introduzindo de forma aleatória valores omissos em cada conjunto de dados de acordo com as percentagens definidas anteriormente. É de notar que antes deste passo foram selecionadas e isoladas as variáveis *target* de cada conjunto para que não sejam afetadas pelo processo, de forma a que posteriormente se possa testar o impacto de cada imputação na classificação da variável.

Depois da implementação dos métodos foram analisados os erros de imputação utilizando NRMSE para as variáveis numéricas e PCMC (percentagem de casos mal classificados) para as variáveis categóricas, comparando os conjuntos imputados com o conjunto original.

8. Resultados

Para cada conjunto de dados será apresentado o erro de imputação na forma gráfica com o intuito de mostrar a sua evolução com o aumento da percentagem de valores omissos. Estes gráficos são relativos aos valores de NRMSE e PCMC, sendo ambos apresentados para todos os conjuntos, com exceção daqueles que possuem apenas um tipo de variáveis, para os quais só será apresentado o gráfico relativo a esse tipo.

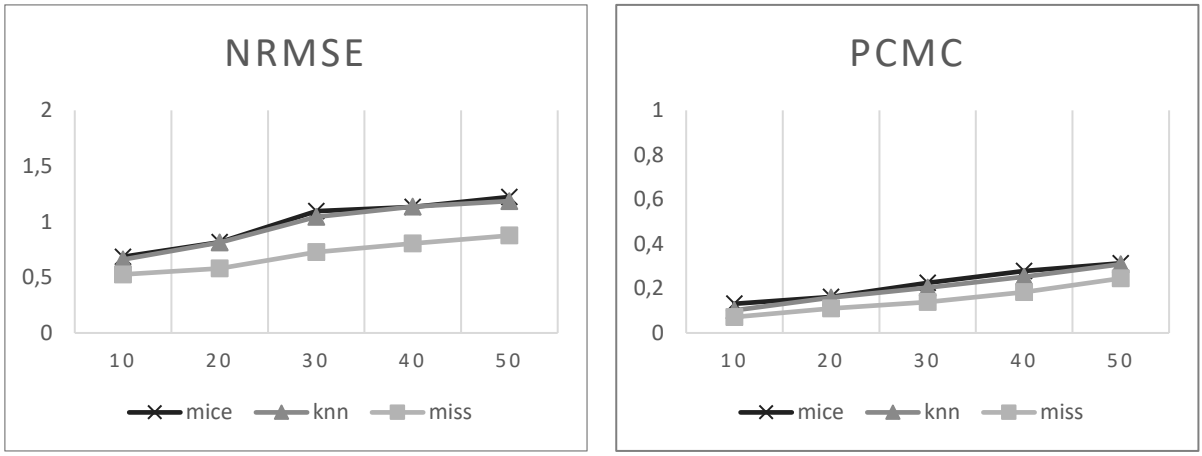


Figura 8.1: NRMSE e PCMC para o dataset Acute Inflammations

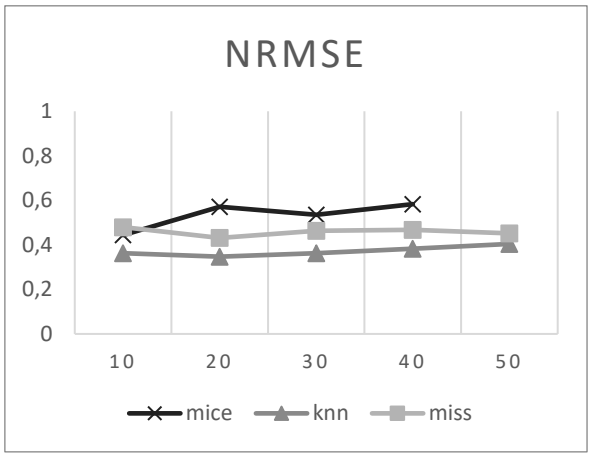


Figura 8.2: NRMSE para o dataset Autistic Spectrum Disorder Screening Data for Adolescent

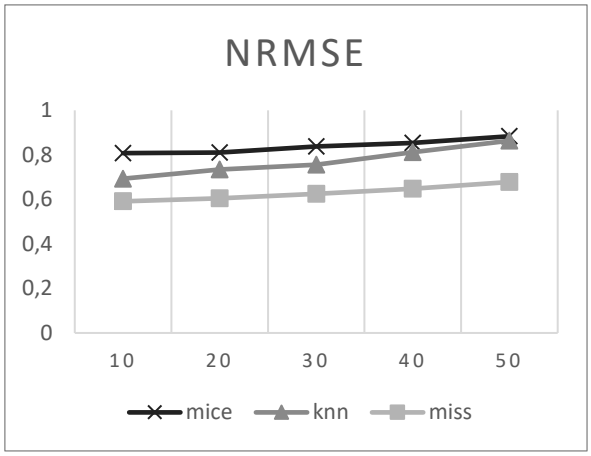
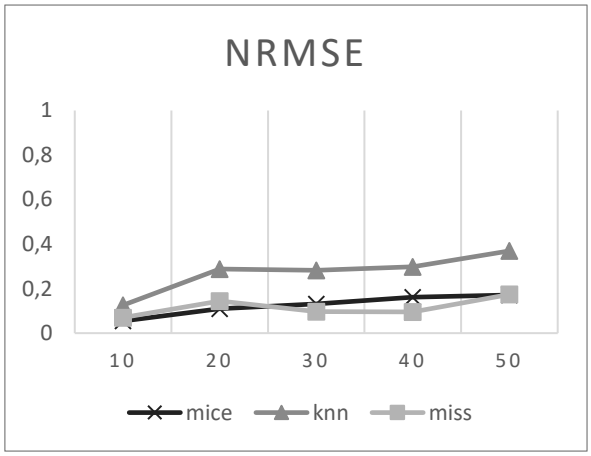


Figura 8.3: NRMSE para o dataset Breast Cancer



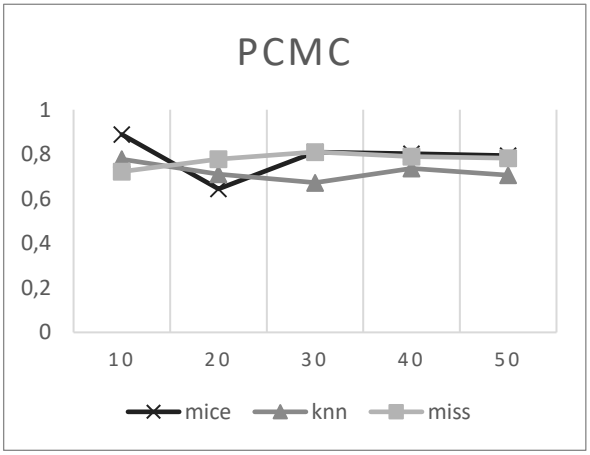


Figura 8.4: NRMSE e PCMC para o dataset Breast Cancer Wisconsin (Prog.)

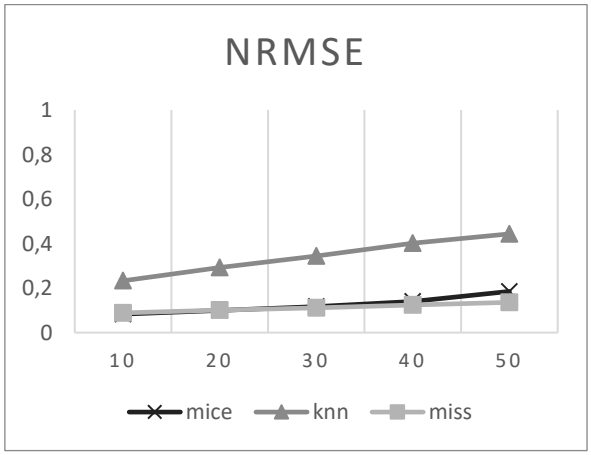


Figura 8.5: NRMSE para o dataset Breast Cancer Wisconsin (Diagnostic)

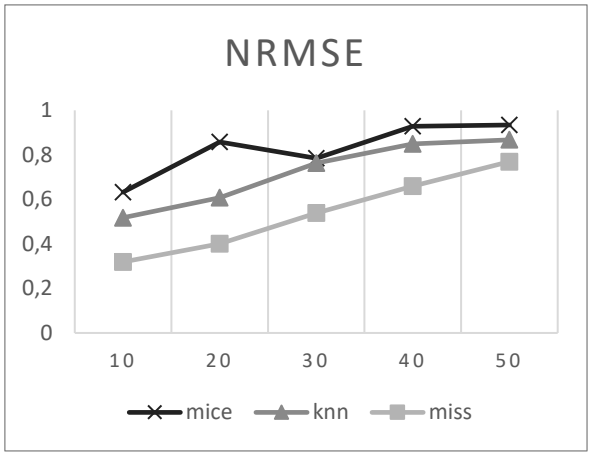
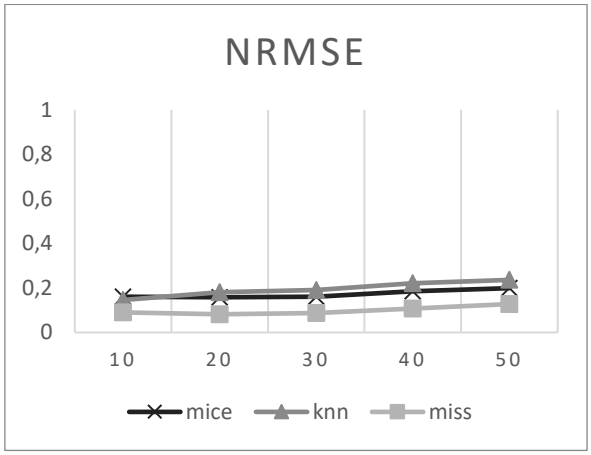


Figura 8.6: NRMSE para o dataset Breast Tissue



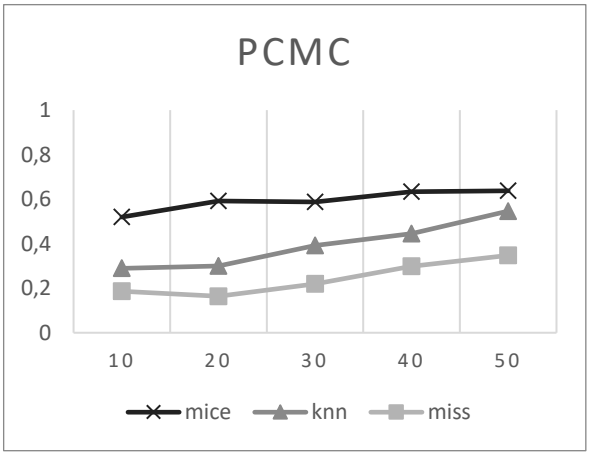
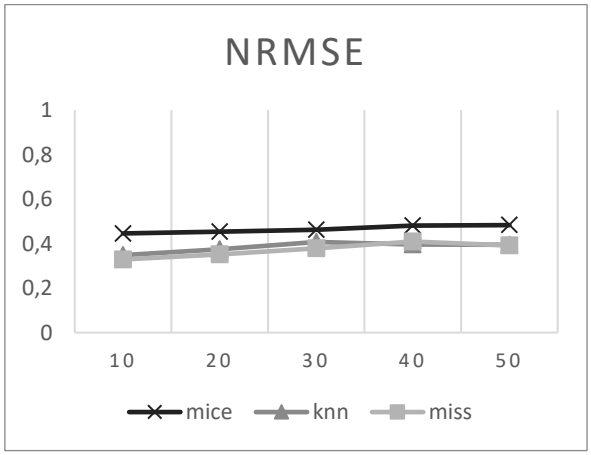


Figura 8.7: NRMSE e PCMC para o dataset Cardiotocography



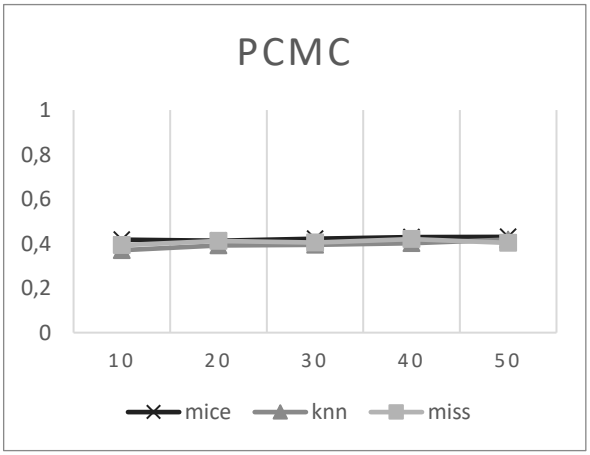


Figura 8.8: NRMSE e PCMC para o dataset Contraceptive Method Choice

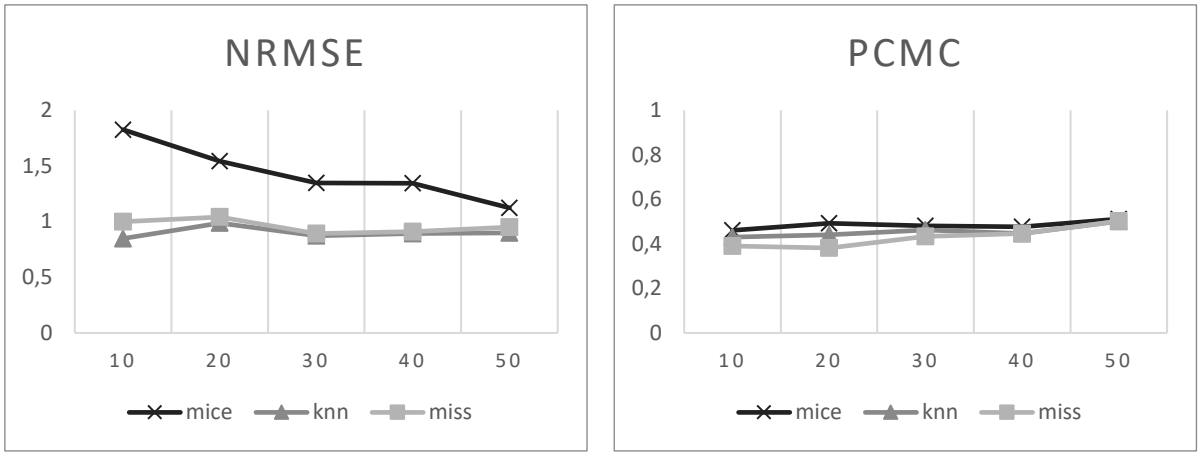


Figura 8.9: NRMSE e PCMC para o dataset Cryotherapy

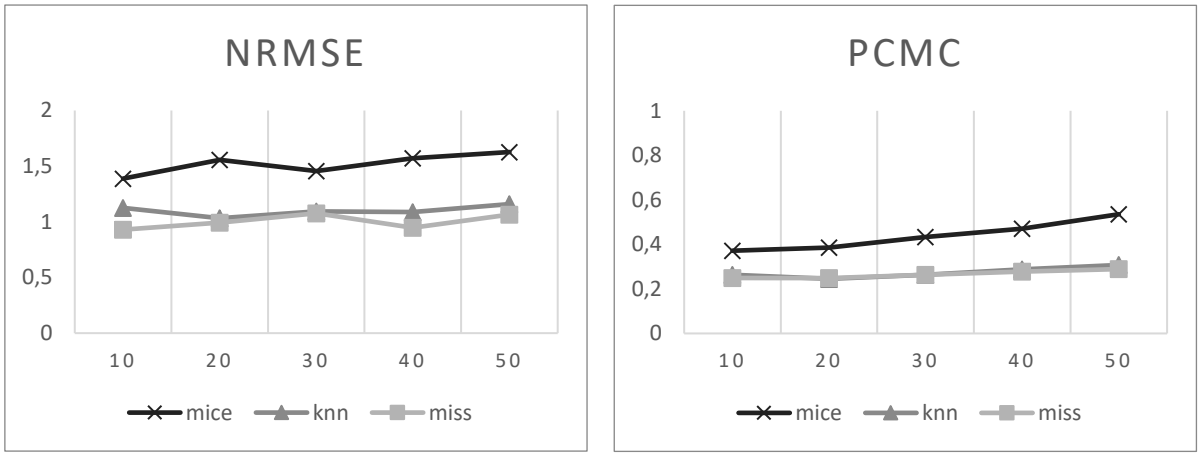


Figura 8.10: NRMSE e PCMC para o dataset Dermatology

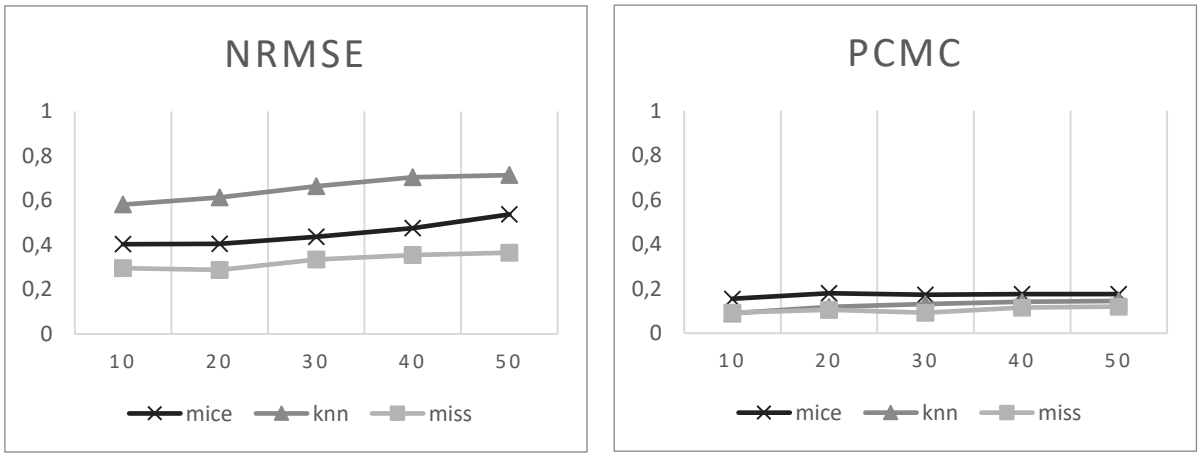


Figura 8.11: NRMSE e PCMC para o dataset Diabetic Retinopathy Debrecen

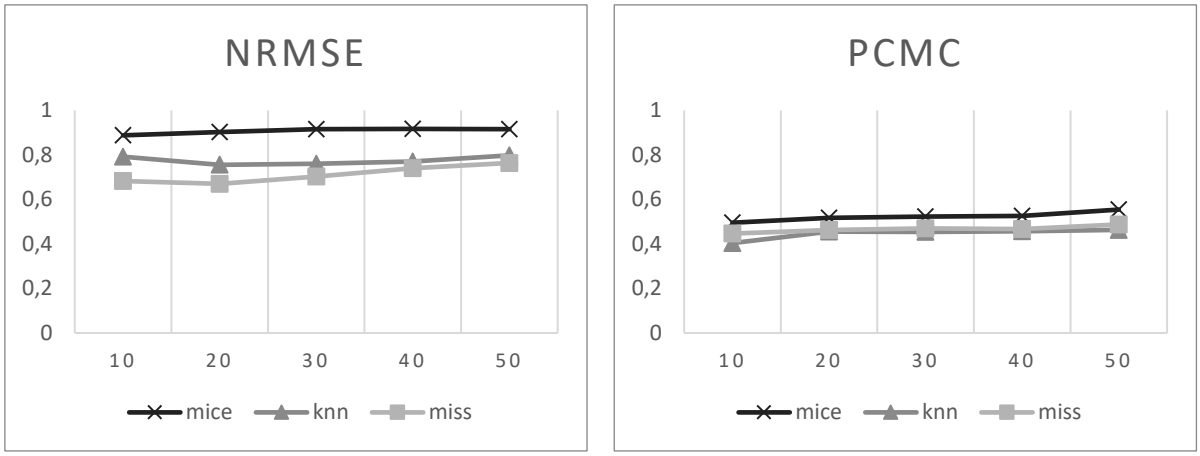


Figura 8.12: NRMSE e PCMC para o dataset Fertility

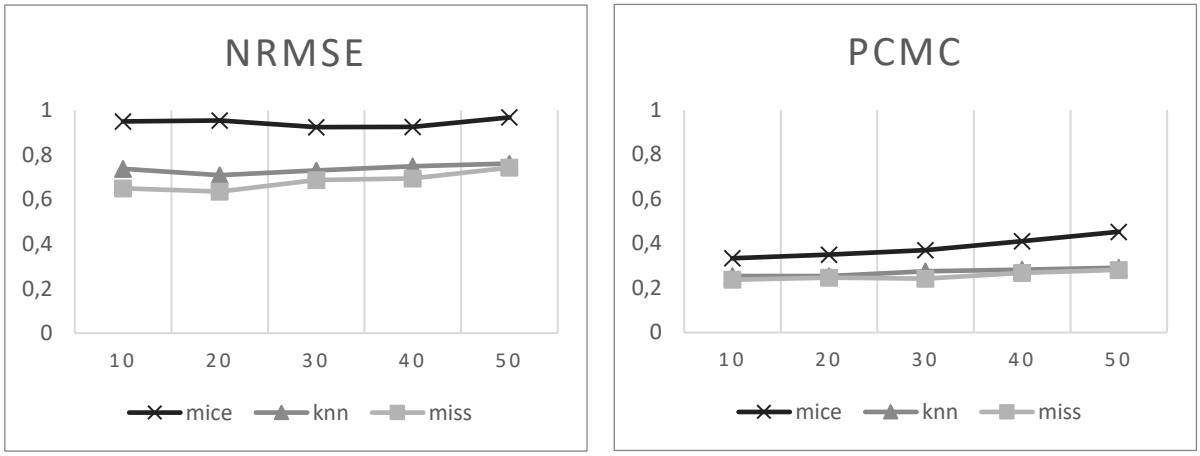


Figura 8.13: NRMSE e PCMC para o dataset Hepatitis

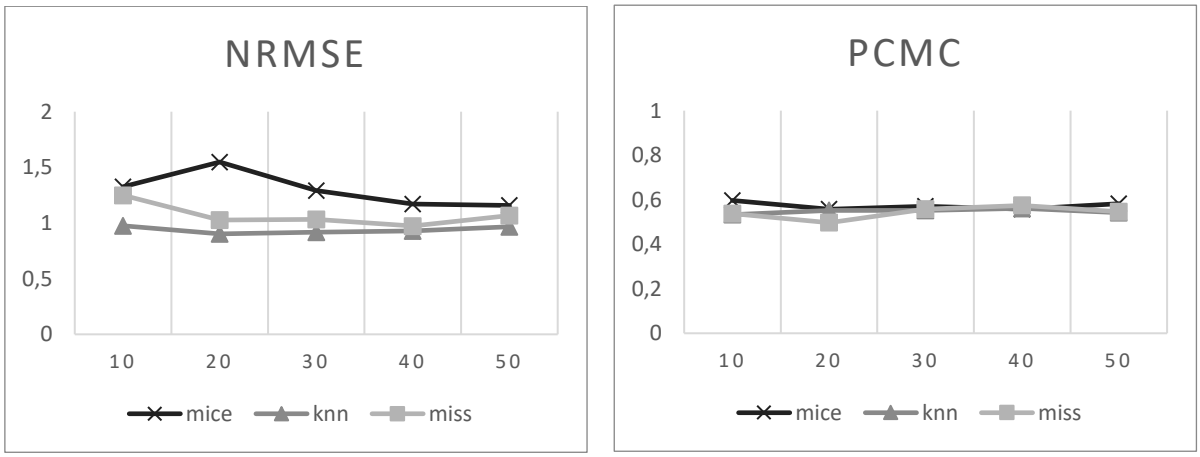


Figura 8.14: NRMSE e PCMC para o dataset Immunotherapy

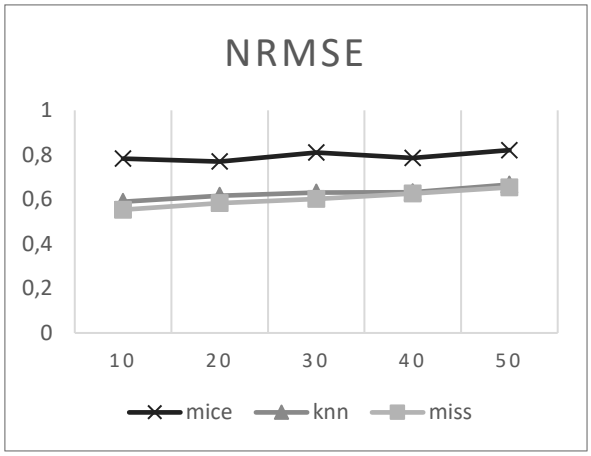


Figura 8.15: NRMSE e PCMC para o dataset Liver Disorders

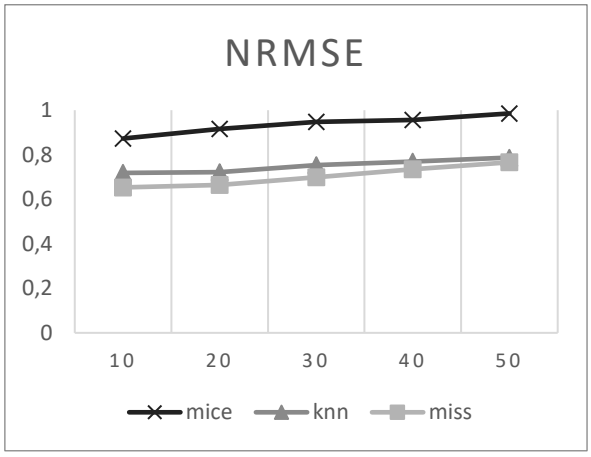
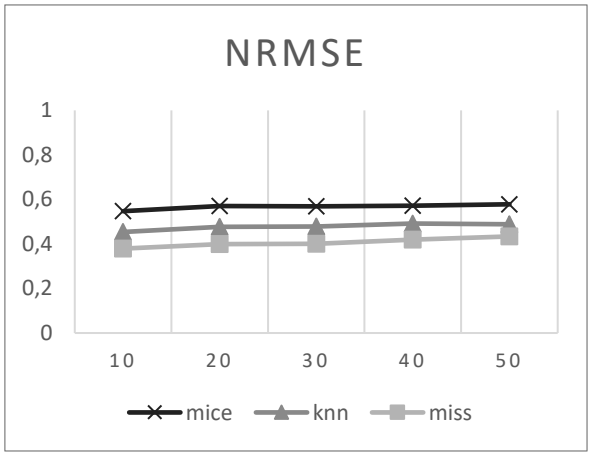


Figura 8.16: NRMSE e PCMC para o dataset Pima Indians Diabetes



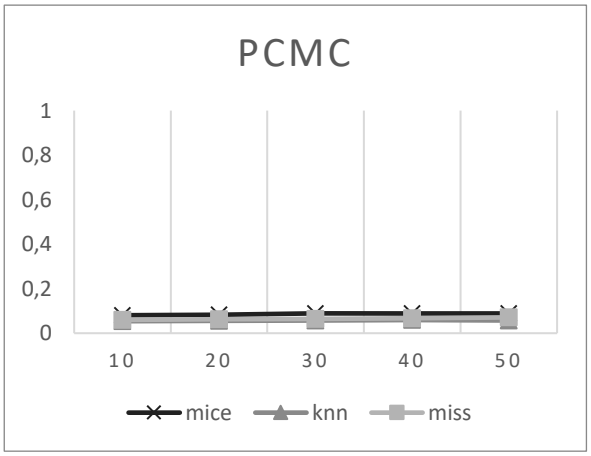


Figura 8.17: NRMSE e PCMC para o dataset Thyroid Disease

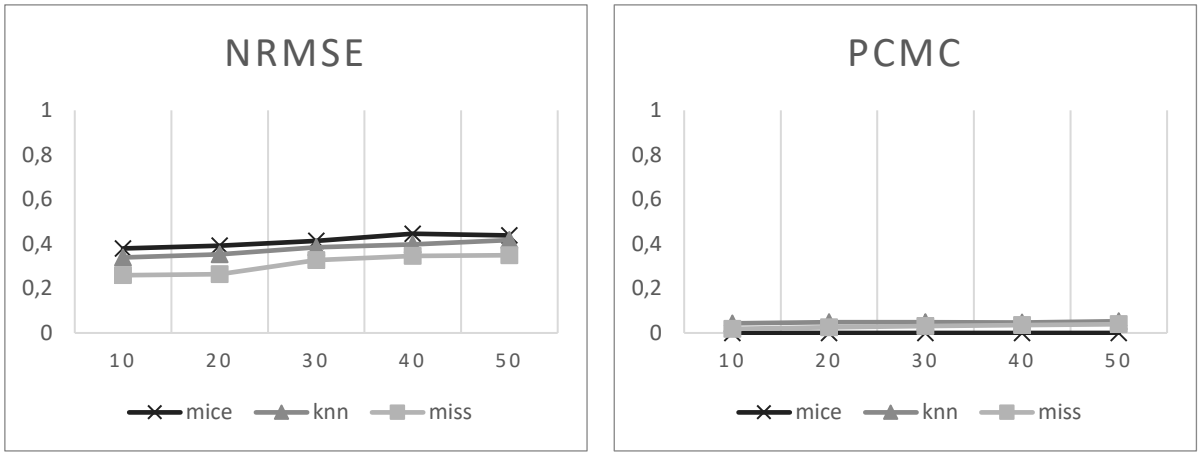


Figura 8.18: NRMSE e PCMC para o dataset Cervical cancer (Risk Factors)

Para além do erro de imputação associado aos métodos, foi também estudado qual o impacto que cada imputação tem na classificação da variável *target* em cada conjunto de dados. Foram assim aplicados dois métodos de classificação (kNN e *Random Forest*) a cada conjunto de dados, originais e imputados, analisado depois a evolução do valor da acurácia.

	MICE		kNN		missForest	
	RF	kNN	RF	kNN	RF	kNN
0%	1,000	1,000	1,000	1,000	1,000	1,000
10%	0,972	0,944	1,000	1,000	1,000	1,000
20%	0,944	0,944	0,972	0,972	1,000	1,000
30%	0,972	0,972	1,000	1,000	0,889	0,889
40%	0,944	0,861	0,917	0,889	0,889	0,861
50%	0,889	0,889	0,889	0,861	0,861	0,833

Tabela 8.1: Acurácia da classificação para o dataset Acute Inflammations

	MICE		kNN		missForest	
	RF	kNN	RF	kNN	RF	kNN
0%	0,964	0,964	0,964	0,964	0,964	0,964
10%	0,929	0,750	0,929	0,857	0,964	0,929
20%	0,786	0,786	0,821	0,821	1,000	0,929
30%	0,929	0,857	1,000	0,857	0,929	0,893
40%	0,714	0,786	0,821	0,821	0,786	0,714
50%	-	-	0,857	0,714	0,857	0,857

Tabela 8.2: Acurácia da classificação para o dataset Autistic Spectrum Disorder Screening Data for Adolescent

	MICE		kNN		missForest	
	RF	kNN	RF	kNN	RF	kNN
0%	0,985	0,985	0,985	0,985	0,985	0,985
10%	0,951	0,956	0,966	0,956	0,956	0,956
20%	0,985	0,985	0,961	0,961	0,961	0,975
30%	0,985	0,980	0,966	0,951	0,961	0,961
40%	0,975	0,975	0,966	0,946	0,971	0,956
50%	0,922	0,931	0,961	0,961	0,966	0,966

Tabela 8.3: Acurácia da classificação para o dataset Breast Cancer

	MICE		kNN		missForest	
	RF	kNN	RF	kNN	RF	kNN
0%	0,772	0,754	0,772	0,754	0,772	0,754
10%	0,772	0,754	0,789	0,754	0,807	0,702
20%	0,772	0,737	0,772	0,684	0,772	0,702
30%	0,772	0,719	0,772	0,737	0,772	0,772
40%	0,807	0,737	0,789	0,754	0,772	0,772
50%	0,772	0,772	0,789	0,772	0,772	0,754

Tabela 8.4: Acurácia da classificação para o dataset Breast Cancer Wisconsin (Prog.)

	MICE		kNN		missForest	
	RF	kNN	RF	kNN	RF	kNN
0%	0,935	0,941	0,935	0,941	0,935	0,941
10%	0,941	0,912	0,953	0,912	0,941	0,924
20%	0,965	0,947	0,947	0,894	0,935	0,918
30%	0,959	0,935	0,959	0,906	0,953	0,935
40%	0,965	0,918	0,953	0,941	0,971	0,941
50%	0,971	0,941	0,929	0,924	0,912	0,900

Tabela 8.5: Acurácia da classificação para o dataset Breast Cancer Wisconsin (Diagnostic)

	MICE		kNN		missForest	
	RF	kNN	RF	kNN	RF	kNN
0%	0.552	0.380	0.552	0.380	0.552	0.380
10%	0,586	0,345	0,655	0,483	0,793	0,448
20%	0,655	0,310	0,690	0,517	0,724	0,655
30%	0,552	0,448	0,690	0,379	0,655	0,483
40%	0,724	0,517	0,517	0,448	0,690	0,483
50%	0,517	0,414	0,483	0,379	0,448	0,414

Tabela 8.6: Acurácia da classificação para o dataset Breast Tissue

	MICE		kNN		missForest	
	RF	kNN	RF	kNN	RF	kNN
0%	0,991	0,885	0,991	0,885	0,991	0,885
10%	0,984	0,882	0,975	0,910	0,992	0,910
20%	0,981	0,885	0,975	0,890	0,973	0,892
30%	0,965	0,896	0,959	0,873	0,967	0,890
40%	0,951	0,903	0,936	0,884	0,934	0,881
50%	0,909	0,862	0,926	0,876	0,936	0,885

Tabela 8.7: Acurácia da classificação para o dataset Cardiotocography

	MICE		kNN		missForest	
	RF	kNN	RF	kNN	RF	kNN
0%	0,500	0,526	0,500	0,526	0,500	0,526
10%	0,518	0,505	0,502	0,518	0,541	0,518
20%	0,461	0,459	0,489	0,459	0,489	0,520
30%	0,464	0,445	0,516	0,482	0,482	0,466
40%	0,452	0,430	0,452	0,434	0,486	0,448
50%	0,418	0,405	0,475	0,443	0,459	0,436

Tabela 8.8: Acurácia da classificação para o dataset Contraceptive Method Choice

	MICE		kNN		missForest	
	RF	kNN	RF	kNN	RF	kNN
0%	0,962	0,846	0,962	0,846	0,962	0,846
10%	0,923	0,808	0,923	0,769	0,885	0,769
20%	0,846	0,577	0,654	0,654	0,808	0,808
30%	0,731	0,808	0,808	0,769	0,885	0,692
40%	0,808	0,615	0,731	0,731	0,731	0,692
50%	0,808	0,692	0,692	0,654	0,615	0,692

Tabela 8.9: Acurácia da classificação para o dataset Cryotherapy

	MICE		kNN		missForest	
	RF	kNN	RF	kNN	RF	kNN
0%	0,781	0,714	0,781	0,714	0,781	0,714
10%	0,781	0,752	0,819	0,724	0,781	0,752
20%	0,790	0,743	0,800	0,714	0,781	0,733
30%	0,781	0,762	0,829	0,714	0,781	0,743
40%	0,762	0,714	0,790	0,771	0,762	0,752
50%	0,790	0,733	0,819	0,762	0,752	0,781

Tabela 8.10: Acurácia da classificação para o dataset Dermatology

	MICE		kNN		missForest	
	RF	kNN	RF	kNN	RF	kNN
0%	0,664	0,623	0,664	0,623	0,664	0,623
10%	0,658	0,661	0,672	0,687	0,716	0,704
20%	0,617	0,641	0,658	0,643	0,643	0,646
30%	0,667	0,649	0,635	0,638	0,638	0,638
40%	0,655	0,641	0,617	0,606	0,646	0,612
50%	0,626	0,603	0,635	0,583	0,635	0,594

Tabela 8.11: Acurácia da classificação para o dataset Diabetic Retinopathy Debrecen

	MICE		kNN		missForest	
	RF	kNN	RF	kNN	RF	kNN
0%	0,931	0,897	0,931	0,897	0,931	0,897
10%	0,897	0,897	0,862	0,897	0,897	0,897
20%	0,897	0,897	0,897	0,897	0,897	0,897
30%	0,897	0,897	0,862	0,897	0,828	0,897
40%	0,897	0,897	0,897	0,897	0,828	0,897
50%	0,862	0,897	0,897	0,897	0,897	0,897

Tabela 8.12: Acurácia da classificação para o dataset Fertility

	MICE		kNN		missForest	
	RF	kNN	RF	kNN	RF	kNN
0%	0,844	0,844	0,844	0,844	0,844	0,844
10%	0,813	0,844	0,875	0,844	0,875	0,781
20%	0,906	0,813	0,906	0,844	0,875	0,844
30%	0,844	0,844	0,875	0,844	0,906	0,844
40%	0,813	0,844	0,844	0,875	0,875	0,844
50%	0,844	0,813	0,813	0,813	0,813	0,844

Tabela 8.13: Acurácia da classificação para o dataset Hepatitis

	MICE		kNN		missForest	
	RF	kNN	RF	kNN	RF	kNN
0%	0,846	0,731	0,846	0,731	0,846	0,731
10%	0,923	0,808	0,769	0,808	0,808	0,808
20%	0,846	0,769	0,846	0,808	0,846	0,808
30%	0,808	0,769	0,846	0,846	0,808	0,808
40%	0,885	0,808	0,808	0,808	0,846	0,769
50%	0,769	0,769	0,846	0,769	0,846	0,808

Tabela 8.14: Acurácia da classificação para o dataset Immunotherapy

	MICE		kNN		missForest	
	RF	kNN	RF	kNN	RF	kNN
0%	0,647	0,627	0,647	0,627	0,647	0,627
10%	0,549	0,529	0,451	0,520	0,490	0,559
20%	0,618	0,520	0,608	0,627	0,539	0,569
30%	0,549	0,471	0,539	0,549	0,510	0,549
40%	0,520	0,500	0,490	0,559	0,559	0,598
50%	0,549	0,471	0,569	0,569	0,539	0,578

Tabela 8.15: Acurácia da classificação para o dataset Liver Disorders

	MICE		kNN		missForest	
	RF	kNN	RF	kNN	RF	kNN
0%	0,765	0,757	0,765	0,757	0,765	0,757
10%	0,761	0,704	0,739	0,761	0,765	0,717
20%	0,743	0,730	0,713	0,674	0,735	0,743
30%	0,722	0,726	0,726	0,700	0,709	0,670
40%	0,730	0,722	0,713	0,687	0,713	0,713
50%	0,748	0,735	0,700	0,696	0,683	0,657

Tabela 8.16: Acurácia da classificação para o dataset Pima Indians Diabetes

	MICE		kNN		missForest	
	RF	kNN	RF	kNN	RF	kNN
0%	0,946	0,942	0,946	0,942	0,946	0,942
10%	0,952	0,939	0,949	0,935	0,947	0,939
20%	0,947	0,934	0,949	0,933	0,944	0,938
30%	0,941	0,933	0,943	0,936	0,941	0,938
40%	0,932	0,932	0,936	0,930	0,941	0,935
50%	0,935	0,933	0,935	0,934	0,937	0,934

Tabela 8.17: Acurácia da classificação para o dataset Thyroid Disease

	MICE		kNN		missForest	
	RF	kNN	RF	kNN	RF	kNN
0%	0,970	0,935	0,970	0,935	0,970	0,935
10%	-	-	0,965	0,930	0,965	0,935
20%	-	-	0,945	0,935	0,960	0,935
30%	-	-	0,970	0,935	0,935	0,935
40%	-	-	0,950	0,935	0,940	0,935
50%	-	-	0,935	0,935	0,935	0,935

Tabela 8.18: Acurácia da classificação para o dataset Cervical cancer (Risk Factors)

Erros de imputação

É facilmente observável que MICE é o que tende a obter a pior performance entre os três métodos utilizados tanto ao nível das variáveis numéricas como ao nível das variáveis categóricas. Apesar de em alguns casos em que a sua capacidade de imputação tem uma boa performance, de uma forma geral kNN e *missForest* obtêm melhores resultados. É de notar que este método por vezes falha a imputação de variáveis categóricas, como foi possível observar para o conjunto de dados 18 (*Cervical cancer (Risk Factors))*, onde o método fez a imputação de variáveis categóricas.

Analisando agora o método kNN temos que este tende a ter um menor erro de imputação que o MICE, no entanto existem casos em que o valor deste erro excede por larga margem qualquer dos outros métodos, sendo que em 3 conjuntos de dados (2, 9 e 14) foi o método com menor erro de imputação.

Por último, o método *missForrest*, é que obtém na maioria dos testes o menor erro, tanto ao nível de variáveis categóricas como numéricas. Mesmo nos casos em que este não foi o melhor método, nunca se afastou do menor erro obtido. Foi desta forma considerado o método mais consistente para fazer a imputação, no entanto esta maior precisão possui um custo, o tempo.

Analisandos os tempos de imputação é possível observar grandes disparidades. O método kNN é aquele que obteve sempre o melhor tempo, o que o torna num método muito aliciante para imputações rápidas, e devido ao facto que a sua performance tende a ser relativamente boa comparando com os outros métodos, é normal que esta metodologia de imputação seja bastante difundida e utilizada.

Comparando os tempos de imputação dos métodos, observou-se que MICE e *missForest* levam sempre mais tempo que kNN, sendo que este tende a demorar entre alguns segundos a poucos minutos, enquanto MICE e *missForest* tendem a demorar entre vários minutos a poucas horas se a dimensão do conjunto de dados for elevada.

O aumento do tempo de imputação destes métodos é, no entanto, afetado de formas diferentes, o MICE tende a sofrer um impacto mais significativo quanto maior for o número de variáveis, ao invés do *missForest* que tende a ser mais impactado pelo aumento das observações.

Depois da análise destes resultados é possível então tirar 3 grandes conclusões:

- O método *missForest* é o mais consistente dos 3 utilizados, sendo que o tempo de imputação pode ser muito longo para grandes conjuntos de dados com muitas observações.
- O método MICE é o menos consistente, sendo que o tempo que demorara a imputar os valores omissos também pode ser bastante longo.
- O método kNN é muito rápido, comparativamente aos outros dois, obtendo por norma resultados equiparáveis, sendo que por vezes consegue obter também o menor erro de imputação.

De forma a avaliar o impacto que a imputação teve na classificação foi comparada a acurácia que cada método obteve, ou seja, a sua taxa de acertos. Esta foi a medida selecionada porque, apesar das conhecidas limitações que esta medida tem para avaliar conjuntos com uma variável *target* não balanceada [13], continua a ser uma medida que permite avaliar de forma geral a *performance* de um modelo.

No conjunto de dados *Autistic Spectrum Disorder Screening Data for Adolescent* não foi possível calcular a acurácia para o conjunto de dados imputado pelo método MICE com 50% de omissos, pois este não consegui imputar todos os valores omissos, como foi possível observar no gráfico de evolução do NRMSE. O mesmo problema aconteceu na imputação do conjunto de dados *Cervical cancer (Risk Factors).*

Foi possível observar que independentemente do método de imputação a taxa de acerto tende a diminuir com o aumento da percentagem de omissos imputados, embora em alguns conjuntos de dados este decréscimo seja mínimo. Em alguns casos observou-se que a taxa de acertos se manteve aproximadamente igual com o aumento de omissos, tendo em raras exceções aumentado.

Comparando os conjuntos de dados imputados pelos três métodos, MICE, kNN e *missForest*, observou-se que, apesar de existirem discrepâncias no impacto dos métodos dentro do mesmo conjunto de dados, esta discrepância não foi generalizada para os restantes conjuntos. Desta forma não houve nenhum método que, de forma geral, causasse menos impacto na classificação que os restantes.

9. Conclusões

A existência de valores omissos em conjuntos de dados é, e provavelmente sempre será um grande problema quer na área clínica como em qualquer outra área, pelo que encontrar as melhores formas de contornar este problema é muito importante.

Neste trabalho comparou-se três metodologias de imputação em R bastante conhecidas, MICE, kNN e *missForest*.

Foram rapidamente sentidas dificuldades na procura de conjuntos de dados completos necessários para este trabalho, pelo que se tentou encontrar os melhores e mais completos conjuntos disponíveis. A procura de dados foi bastante facilitada pelo repositório UCI *Machine Learning Repository*, repositório que disponibiliza uma grande variedade de conjuntos de dados de várias áreas incluindo a clínica.

Após a obtenção e pré-tratamento de todos os conjuntos de dados necessários para este trabalho foi necessário desenvolver algoritmos que permitissem obter os resultados da forma mais eficaz possível. Fóruns *on-line* como *Stackoverflow* e *RBloggers* foram uma grande ajuda no desenvolvimento destes algoritmos.

Analisando então os resultados obtidos podemos concluir que a metodologia *missForest* é a que apresenta o menor erro de imputação na maioria dos casos, sendo que MICE foi a que teve tendência a gerar o maior erro.

A metodologia kNN, apesar da sua simplicidade, obteve resultados satisfatórios superando várias vezes MICE, e por vezes a *missForest*. Este método é, no entanto, aquele que menos tempo demorar a imputar os dados, sendo que os outros métodos tendem a demorar longos períodos de tempo ao imputar valores omissos em conjuntos de dados de maior dimensão.

Os resultados do presente estudo mostram que o método que na generalidade apresentou o menor erro de imputação foi o *missForest*, o que vai de encontro aos resultados já apresentados por Daniel J. Stekhoven e Peter Bühlmann [12]. Portanto conclui-se que de entre os três métodos comparados neste trabalho o recomendado para obter os melhores resultados de imputação é o *missForest*.

Após os erros de imputação, foi analisado o impacto que a imputação de valores omisso tem na classificação. Observou-se que com o aumento da taxa de omissos imputados a acurácia tendia a descer, o que que vai ao encontro dos resultados obtidos por Youting Sun, Ulisses Braga-Neto e EdwardR Dougherty [14], que também observaram que os resultados da classificação eram geralmente melhores quando se lidava com os conjuntos de dados originais.

No entanto, em raras exceções, a imputação teve um impacto positivo na classificação, isto pode ter acontecido devido à existência de ruído ou pequena variância dos dados [14]. Resultados obtidos por Alireza Farhangfara, Lukasz Kurganb e Jennifer Dy [15], também sugerem que determinados métodos de imputação podem ter um impacto positivo na classificação dos conjuntos de dados.

Referências

[1] Allison, P.D. Missing Data. Thousand Oaks, CA: Sage. 2001;

[2] Bhaskaran, K., & Smeeth, L. What is the difference between missing completely at random and missing at random?. International journal of epidemiology. 2014; 43(4), 1336-9.

[3] Gelman, A., & Hill, J. Missing-data imputation. In Data Analysis Using Regression and Multilevel/Hierarchical Models (Analytical Methods for Social Research). 2006; 529-44.

[4] Kang H. The prevention and handling of the missing data. Korean J Anesthesiol. 2013;64(5):402-6.

[5] Zhang Z. Missing data imputation: focusing on single imputation. Ann Transl Med. 2016; 4(1):9.

[6] Li P, Stuart EA, Allison DB. Multiple Imputation: A Flexible Tool for Handling Missing Data. JAMA. 2015;314(18):1966-7.

[7] Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work?. Int J Methods Psychiatr Res. 2011;20(1):40-9.

[8] Prasath, Surya & Arafat Abu Alfeilat, Haneen & Lasassmeh, Omar & Hassanat, Ahmad. Distance and Similarity Measures Effect on the Performance of K-Nearest Neighbor Classifier - A Review. 2017.

[9] Faisal S., Tutz G. Nearest Neighbor Imputation for Categorical Data by Weighting of Attributes. 2017;

[10] Hu LY, Huang MW, Ke SW, Tsai CF. The distance function effect on k-nearest neighbor classification for medical datasets. Springerplus. 2016;5(1):1304.

[11] Chen, Jiahua & Shao, J.. Nearest neighbor imputation for survey data. Journal of Official Statistics. 2000; 16. 113-131.

[12] Daniel J. Stekhoven, Peter Bühlmann; MissForest—non-parametric missing value imputation for mixed-type data, Bioinformatics. 2008; 28(1): 112–118

[13] Jeni LA, Cohn JF, De La Torre F. Facing Imbalanced Data Recommendations for the Use of Performance Metrics. Int Conf Affect Comput Intell Interact Workshops. 2013:245-251.

[14] Sun Y, Braga-Neto U, Dougherty ER. Impact of missing value imputation on classification for DNA microarray gene expression data--a model-based study. EURASIP J Bioinform Syst Biol. 2010;2009(1):504069.

[15] Farhangfar, A., Kurgan, L., & Dy, J. Impact of imputation of missing values on classification error for discrete data. Pattern Recognition. 2008; 41(12), 3692–3705.

[16] Zhang Z. Introduction to machine learning: k-nearest neighbors. Ann Transl Med. 2016;4(11):218.

[17] Quinlan, J. R., . Induction of Decision Trees. Machine Learning 1, Kluwer Academic Publishers. 1986; 81-106.

[18] Ho, Tin Kam. Random Decision Forests. Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995; 278–282